TRANSFORMING MESSAGE DETECTION

Liana Ermakova
Perm State University
e-mail: liana87@mail.ru

Abstract

The majority of existing spam filtering techniques suffers from several serious disadvantages. Some of them provide many false positives. The others are suitable only for email filtering and may not be used in IM and social networks. Therefore content methods seem to be more efficient. One of them is based on signature retrieval. However it is not change resistant. There are enhancements (e.g. checksums) but they are extremely time and resource consuming. That is why the main objective of this research is to develop a transforming message detection method. To this end we have compared spam in various languages, namely English, French, Russian and Italian. For each language the number of examined messages including spam and notspam was about 1000. 135 quantitative features have been retrieved. Almost all these features do not depend on the language. They underlie the first step of the algorithm based on support vector machine. The next stage is to test the obtained results applying N-gram approach. Special attention is paid to word distortion and text alteration. The obtaining results indicate the efficiency of the suggested approach.

Keywords: *spam, transforming message, n-grams, SVM, Damerau-Levenshtein distance.*

1. INTRODUCTION

Kaspersky Lab defines spam in the following way:

Spam is unsolicited anonymous mass email[1].

According to Kaspersky lab, in the last quarter of 2010 spam made up 77.1% of total email traffic[2]. It should also be mentioned that Russian spam became more carefully designed: more spam messages have the HTML format[3]. Nowadays spam concerns not only email, but also social networks, instant messaging (IM) and other systems. Traditional approaches such as blacklisting and message header analysis are efficient enough for email filtering. Though, they fail to deal with spam in social networks, IM and forums. In this case content and link analyses seem to be more effective.

Spam appeared in the nineties of the XX century. Firstly, spam was sent from proper spammers' addresses. The earliest messages were similar. That spam is easy to filter. Content analysis development forced spam to evolve. All messages became different. One of the ways to do it is to add an address to the beginning of a letter (e.g. adding «Hello, joe!» to the message to joe@user.com). The trick may be detected by applying fuzzy signature or statistical learning filters (like Bayesian filtering). A message may begin or end with an extract from classical literature or a sequence of random

words. HTML message may contain an unreadable text (e.g. printed in very small font or the same color as the background). These additions provide obstacles to fuzzy signature and statistical filters. In response new techniques appear such as quotation searching and detailed HTML parsing. Usually it is possible to detect spammer's trick it-self and classify a message as spam without detailed content analysis. An advertisement may be sent as a picture. Therefore image analysis techniques which enable to retrieve a text from a picture are used [4].

Transforming messages are messages which have the same meaning but different forms. Every message looks like a connected text. Only if one has a number of these letters it is possible to establish a paraphrasing fact.

2. STATE OF THE ART

Perhaps, the most famous spam filter is SpamAssassin (SA). SA has an extensible system of weighted spam detection rules.The algorithm is based on an off-line artificial neural net and a Bayes trainer. It can run the entire rule set directly or accelerate some body rules but not any other rules, such as header, URI or plugin rules. There are many attempts to improve SA, e.g. to translate regex rules into the POSIX format[5]. Regexes are very time consuming. If SA is overtrained the quality goes down[6]. SA fails to deal with untypical mails and may provide false positive results[7]. Moreover it raises a lot of difficulties to Russian users[8].

Some methods are based on TCP fingerprints[9]. The major drawback of this approach seems to be the fact that nowadays the major part of junk emails is delivered fromcompromised user machines, therefore legal messages from these users will be lost. Sometimes spammers try to cheat users and filters by changing addresses and locations. Header analyses are efficient enough in filtering that kind of spam [10]. However it is applicable only to email filtering since it uses housekeeping information which is not available in other spam types. The most widely used tricks are transforming messages, spam sent as a graphic attachment and unreadable text addition. And not all spam filters can deal with them[11].

Yandex divides spam detections methods into two categories:

- Techniques based on text samples (it is difficult to make them and to keep them up to date);
- Manual analysis and email monitoring (e.g. signature approach [12].

Yandex uses,inter alia, white listing [12]. This approach suffers from some serious disadvantages. In the systems with authorization mechanism it is not so easy to send a message to a user for the first time. Moreover, the practice indicates that white listing is not efficient in IM (e.g. qip,

icq) and social networks (e.g. ВКонтакте, Facebook) as far as there the lager half of spam is distributed from the accounts of authorized people. Some researchers believe that spam may be filtered only by end user[13]. According to another survey conducted by Yandex, 40% of the respondents have difficulties in distinguishing spam from legal mail[14].

Today the improvement of signature methods seems to be crucial. There are two basic approaches:

- Syntactical (i.e. operating with word chains);
- Lexical (i.e. operating with dictionary) (e.g. key words)[12].

In current syntactical methods based on shingles (i.e. contiguous subsequences of tokens in a document)[15][16], for each shingle a check sum is computed and then a random sample is constructed from this set. Shingles make it possible to find similar texts rather reliably. However, real-world problems, such as spam filtering, require too many shingles and consequently too many resources in order to cluster messages[12].

The major drawback of every lexical method is that it may be applied only to a single language.

In information retrieval there exists similar problem, namely spamdexing. Spamdexing (or search spam) is the deliberate manipulation of search engine indexes[17].

Search spam may be divided into two parts: link spam and content spam[18]. There are several methods of spamdexing related to content or link spam. The term content spam may be applied to keyword stuffing, hidden or invisible text, meta-tag stuffing, doorway pages, scraper sites or article spinning.

Link spam presupposes such techniques as link-building software, link farms, usage of hidden links, sybil attack, spam blogs and page hijacking[18]. Modern search engines can deal with some kinds of manipulation e.g. applying keyword density. Some algorithms cope with link spam[19]. However one of the most difficult to detect and at the same time one of the most interesting approaches from the linguistic point of view is article spinning. Article spinning is a process of existing articles rewriting in order to avoid penalties for duplicate content. This process is may be accomplished by humans (rewriters or copywriters) or automatically (generated texts). Usually texts are generated on the basis of thesauruses. Sometimes these two methods are integrated: humans write different parts of articles, then special algorithms generates texts which are very closed by meaning but are different in form. Nowadays there are even special instructions how to make various content for hundreds pages (e.g. http://www.seozavr.ru/templates.

html). A half-finished text is similar to the Example 1 (http://bluelinkseo.com/ultra-spinnable-minis/). The same company guaranty that "articles are written only by verified native English speaking writers". It is clear that total number of articles that can be generated *is number of variants of the 1 part x number of variants of the 2 part x number of variants of the 3 part…* So, if there are 3 options for 10 text parts the total number of different articles is 3^{10}=59 049. Another technique is to propose a subject to humans who should produce relatively short texts in a natural language.

Search engines use modifications of the algorithm based on TF×IDF metrics, where TF is a term frequency, that is to say a normalized number of occurrences of a term in the document, and IDF is an inverted document frequency, i.e. a number of documents where this term occurs. Therefore it is possible to manipulate search engine's ranking results. Spammers may try to make one page relevant to many queries by including numerous different terms into the document, or they can improve document relevance by increasing the number of key words[18]. However modern search engines can detect this kind of spam and reduce the rank of respective pages if the density of key words is greater than the normal rate [20]. Usually search engines assign more weight to terms occurring in the tag title than to words appearing in the meta-tags [18]. Key words may occur in hyperlink anchors. In this case the key word is attached to the source page as well as to the linked page. Sometimes spammers just copy somebody else's content, e.g. news [18]. However the majority of search engines try not to index sites that copy others' information and do not produce their own content [21]. Text may be taken partially, e.g. some phrases [18]. Spammers may also use such trick as hidden or invisible text. Another widely spread spamdexing technique is doorway. Doorways (portal pages, jump pages, gateway pages, entry pages) are pages containing key words, meant only for search engines and redirecting users to other pages. The easiest way to do it is to create distinct pages for each query. The next step is to create doorways that are indexed by search engines but do not appear in a browser. The third method is cloaking. In this case users and search engines see different content [22]. Search engines are able to recognize redirect and usually index only the page where a user is redirected [23]. One of the black optimization method is code swapping, i.e. replacement or radical change of the content when the desired position is achieved.

Nowadays automatically generated content detection methods are being developed. The first techniques of search spam filtration were based on heuristics, notably hyperlink analysis, DNS identification, automatic

documents clustering and similar pages detection [24]. Statistical methods were proposed, e.g. number of words on a page (this way is efficient if a spammer tries to increase query covering), number of words in a title, average word length (this method is targeted at queries where key words are not separated by spaces), hyperlink anchor analysis (aimed to identify pages created to increase the ranks of other pages), ratio of visible text to whole content, computation of page redundancy on the basis of shingles or compression ratio, share of stop words, conditional and unconditional n-gram similarity, part of speech analysis (Corleon), General Inquirer (texts are classified to 182 groups) [24]. Methods based on comparison of languages models came from machine translation, e.g. probability of string occurrence in a language [25]. If some pages are marked as spam it is possible to extract similar pages and most likely they are spam. Besides linguistic approaches analysis of pages design is also performed. This method presupposes that spam pages are generated by same patterns [26].

At the present time Markov chain text generation techniques are widely spread. Since the generators are trained at samples of texts in a natural language, methods based on "unnaturalness" of string sequences are not applicable. However contemporary Markov chain text generators are not able to provide full text coherence and stylistic uniformity and this fact may be used for spam filtering [27].

Many search engines are capable to identify non-unique content [21]. Therefore recently new kinds of search engine optimization appeared, namely rewriting and copywriting (article spinning).

There are two main algorithms of page ranking which is based on link analysis: HITS and PageRank [18].

HITS algorithm (Hyperlink-Induced Topic Search, also known as Hubs and authorities) developed by Jon Kleinberg ranks pages according to topics[28]. The first step is to retrieve the set of results to the search query. The computation is performed only on this result set, not across all Web pages. An authority value is computed as the sum of the scaled hub values that point to that page. A hub value is the sum of the scaled authority values of the pages it points to [28]. Hub rank is easy to increase by adding links to "good" pages. Authority rank may be augmented by links with high hub rank. Moreover it is possible to add links to blogs, forums etc., i.e. pages content of which may be modified by a spammer [18].

PageRank assumes that the number of links to a page shows the popularity of this page among users [29]. In order to manipulate this algorithm spammers use link farms. In this case, web pages contain links to the other

pages within a group. Usually these sites do not have useful content. They are created to increase citation index of other pages. Users may be attracted to these site by the content copied from somebody else's pages [18].

A page may be considered as a spam page if it has many of external links and few internal ones [23].

Nowadays there exist a lot of companies which specialize in white or black search engine optimization. Advertisements of purchase and sale of links from the pages with relevant content are frequently found. Moreover it is possible to generate pages from templates, make articles unique or simply page batch loading.

Modern algorithms of link spam detection are based on such indicators as certain words near a link, presence of full link blocks, links to advertising brokers, subject dissimilarity between a link and page content or other links, link position analysis, frequent link updates without content update etc.[19].

3. PECULIARITIES OF SPAM IN VARIOUS LANGUAGES

Spam classification may be made in terms of two criteria: by structure and by subject. Spam may be divided by structure into three types:

- Spam disguising as legal mass mailing;
- Spam disguising as a personal message;
- Advertising spam.

Regardless of the language, advertisement is spam dominating subject, especially medicine, tourism and education offers. English courses are very popular in non-English-speaking countries. Other subjects such as cheap software and pornography are common for various countries.

Advertising spam disguises as legal mass mailing and contains many links (especially French spam) and words related to a commerce sector. It often begins with an exclamatory or interrogative sentence. Bulleted and numbered lists are also common features of spam in various languages. Nevertheless these features may not be used for spam filtering since they occur in legal mass mails.

Another popular subject is easy money (Internet casino, lottery and so on). Sometimes it is related to phishing and identity theft as well as Nigerian scam. The latter resembles personal mail and is difficult to be filtered. Nigerian scam in French is designed according to the rules of business correspondence. However official letters usually contain anexpression «à l'attention de» with a position and/or a name, while in spam one can see «à votre attention». There are a lot of email addresses in business correspondence as well as in phishing. The fraud is that a user may respond to a spam message. In this case the spammer will know that the email is active. The share of spam disguising personal messages is comparatively small. However it is necessary to take them seriously because legal messages can be lost.

French spam is more carefully designed than English and especially Russian ones. Usually it has HTML format therefore there are phrases like "Si ce mailling ne s'affiche pas correctement". Sometimes spammers suggest unsubscribing ("Cliquez ici pour ne plus recevoir nos emails"). If a person clicks on this link the spammer will know that this e-mail address is active and as a result the person will receive more spam or even download a virus. Sometimes spammers "explain" why people receive spam ("Vous êtes inscrit sur", "You are receiving this message because"). Due to perception peculiarity verb forms such as imperative, future simple and present are widespread in spam unlike solicited messages. Direct Impératif is not enough polite. Spammers try to control readers and that is why imperative usually occurs in the aim of a junk mail («push the button now», «achetez maintenant»). An action in indicative mood is thought as a real one.

Many Anglicisms can be found in French and Russian spam. French spam contains less pronouns and possessive determinative than legal messages. There is not such a tendency in the Russian and English languages.

All types of spam appeal to feelings (curiosity, covetousness, laziness, credulity, boredom etc.). Spam features may appear according to subject, structure or aim of a message.

4. METHODS OF MESSAGE TRANSFORMING

Transliteration is often used in Russian spam. Besides, there are a lot of deliberate word distortions (e.g. unnecessary symbols, deliberate misprints, Latin letters in Cyrillic text etc.). However these features do not definitely indicates spam. Sometimes transliteration is used by emigrants and travellers for lack of Russian keyboard layout. Encoding problems may appear. People often apply different transliteration rules. In this case a human being may easily read a message but it is difficult to perform an automated analysis.

Spam	Not Spam
pRODAVA email BAZ pRODAVA BAZ email ADRESOW (ADRESA DLQ email RASSYLOK) eSLIwY OBLADAETE SOBSTWENNYMI INSTRUMENTAMI PROWEDENIQ email RASSYLOK, TO DLQ wAS MY MOVEM PREDLOVITX BAZY DANNYH SOBSTWENNOGO SBORA. <…>cENA ZA 1 MLN. - 50$ cENA ZA WS@ BAZU - 500$ <…>PO L@BYM WOPROSAM: tELEFON:	Privet,zolotze. Nakonez-to posylayu tebe fotki. Ya vybrala nemnozhko bolshe, chtoby ty vybrala kakie hochesh i posmeyalas nemnozhko. Ya kogda smotrela, u menya srazu podnyalos nastroenie. Vse-taki my klassno s toboj syezdili v Ust-Kachku. Esli hochesh, ya tebe vse ostalnye tozhe pereshlu. Pishu tebe iz doma pervyj raz. Ladno, pobezhala delat chto-nibud. A-to zeloe utro za kompiuterom sizhu. Lublu, zeluyu. Mame i koshkam privet!

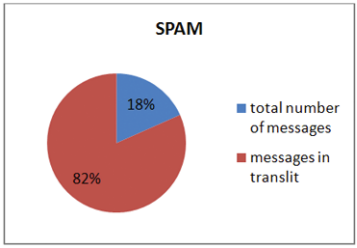

Figure 1. Share of letter written in transliteration in spam

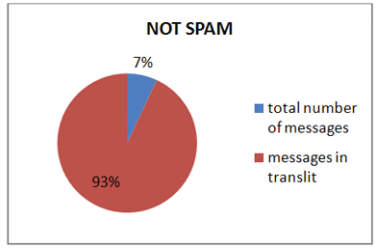

Figure 2. Share of letter written in transliteration in legal messages

Here are some examples of transforming messages written in transliteration.

sWEVIE email BA-ZYpRODAVA BAZ email ADRESOW (ADRESA DLQ email RASSYLOK) <...>	В начале года всегда возникает потребность в “свежих” выписках ЕГРЮЛ и справках Госкомстата. Предлагаем Вам: получение выписки ЕГРЮЛ за 1,2ООрублей справки Госкомстата за 1 200 руб. заказ выписки ЕГРЮЛ + справки Госкомстата составит всего 2.0ОО рублей Доставка курьером, оплата по факту. Контактнаяинформация + 7495 222+07.68
sWEVIE email BA-ZYpRODAVA BAZ email ADRESOW (ADRESA DLQ email RASSYLOK) <...>	В начале года всегда возникает необходимость в “свежих” выписках ЕГРЮЛ и справках Госкомстата. Мы предлагаем Вам: получение выписки ЕГРЮЛ за 1 2ООрублей справки Госкомстата за 1 тыс. 200 р. заказ выписки ЕГРЮЛ + справки Госкомстата составит всего 2 тыс. 0ОО руб-й.Доставкакурьером, оплатапофакту. Телефон: + 7495 222_07;68
aDRESA DLQ email RASSYLOK pRODAVA BAZ email ADRESOW (ADRESA DLQ email RASSYLOK) <...>	В начале года всегда возникает потребность в “свежих” выписках ЕГРЮЛ и справках Госкомстата. Мы предлагаем Вам: получение выписки ЕГРЮЛ за 1 тыс. 200 руб-й справки Госкомстата за 1 200 рублей.заказ выписки ЕГРЮЛ + справки Госкомстата составит всего 2,000 р. Доставка курьером, оплата по факту. Контакты + 7(495) 222-07-68

In spam one can find a lot of “spammers’ tricks”:

- Word distortion;
- Synonymous substitution;
- Substitution of letters by digits and vice versa (4-ч, 0-о, 3-з, 1-l);
- Substitution of Cyrillic symbols by similar Latin letters (k-к, a-a, H –Н ит.д.);
- Unnecessarysymbolsandblanks («Вы хотите вернуть вашего любимо гочелове ка навсегда и полность ю избавиться от измен?»);
- Interchanging of different symbols (e.g., in telephone number).

It is important to mention another transformation technique, namely synonymous expressions (sWEVIE email BAZY = sWEVIE email BAZY = aDRESA DLQ email RASSYLOK, ПредлагаемВам = МыпредлагаемВам, необходимость= потребность).

It happens that only an address or a link transforms:

<...> La preghiamo di rispondere solo alla mia personale e-mail:khhaykanush@yahoo.com Tua amica Haykanush.
<...>La preghiamo di rispondere solo alla mia personale e-mail:haykanusharm@yahoo.com Tua amica Haykanush.
<...>La preghiamo di rispondere solo alla mia personale e-mail:khaykanush@yahoo.com Tua amica Haykanush.

Medicine advertisement is the most changeable. Both a subject and a text transform. Theymayevensubstituteeachother. Usually all links are different (they are automatically created in free hosts). Meanwhile sense is the same.

Subject	Text
Desire to impress and please your lover tonight	The only bluepill you need to get bigger python. http://wanzulkifli.com/c6ave6lc.html
Gain in size and win your wife's addiction	Desire to act like a pornstar? Bang a magicpilule! http://bpyasociados.com.ar/9vh6w3lf.html
Wish to act like a porn-director Nail a blu colored med!	0% amorous failure risk http://mikloswowmobile.com/uaagzeib.html
Dream to act like a porn-director Bang a blu colored pil!	Long manliness is great http://antalyagunlugu.com/d4zz8qan.html

The same can be said about casino. It should be noticed that French and English spam is more intricate than Russian and Italian one; especially it concerns such areas as casino, medicine, stock market games, porno and software. In Spanish there are almost no transformations.

Subject	Text
Comme Faire _200 de _20 - nous APPRENDRONS	Bonne journee Jessikaparsons, {http://yxaqih983.o-f.com/kerizev.html} Accueillez la fortune dans votre vie avec de grandes opportunitйs de gagner, avec l'assurance que vos informations personnelles sont protйgйes et vos gains seront payйs rapidement. Une demi-heure et Ђ200 dans ta poche
Gagner _100 pour une demi-heure c'est rйel	Du jour reussi Shirley_patel, {http://gamingworldshop.ru} Il y a de grandes promotions auxquelles vous pouvez participer et qui vous promettent encore plus de plaisirs et de faзons de gagner. Faire Ђ100 pour une demi-heure - Apprendre?

Subject	Text
Faire -100 pour une demi-heure - Apprendre	Bonne journee Nvshamshik, {http://beluwulod.maddsites.com/abimogek.html} Il y a de grandes promotions auxquelles vous pouvez participer et qui vous promettent encore plus de plaisirs et de faГons de gagner. Gagner -100 pour une demi-heurec'estrИel
Jouer ici, c'est le bonheur ! Telechargez maintenant	{http://opakypiwel.dreamstation.com/jededila.html} On ne peut pas faire plus simple, il suffit de vous inscrire, de faire un versement et vous recevez un fantastique bonus de bienvenue - alors foncez et gagnez ! La meilleure selection de jeu sur internet ! Jouez ici
Jouez plus, gagnez plus	Salut Shea.swan Des options bancaires sыres qui conviendront a tous sont disponibles. Relaxez-vous et soyez certains que vos informations confidentielles sont sйcurisйes et ne seront pas divulguйes. {http://durl.me/554k6}Comment aimeriez-vous commencer au mieux dans le jeu en ligne avec 1,200 Gratuits? Ils sont dйjaa vous, rйclamez-les, jouez et gagnez!

5. N-GRAMS IN TRANSFORMING MESSAGE DETECTION

There are many approaches to find the distance between two documents (e.g. Jaccard coefficient, Hamming distance, edit distance) [30]. In this research we have used N-gram distance.

Traditionally N-grams are used in problems of plagiarism detection [31,32] and language and encoding identification[33,34]. Another group of affiliation methods is based on quantitative text characteristics [35,36,37]. Firstly quantitative features were used in Flesch index and Flesch-Kincaid Index [38]. Within the bound of this work these two approaches have been combined.

We have compared spam in various languages, namely English, French, Russian and Italian. For each language the number of examined messages including spam and notspam was about 1000. Each language sample contains transforming messages but they differ by their subjects. The number of transforming messages of various topics is also different.

We have used 135 quantitative text features such as share of content and function words, share of sentences, paragraph sand words of specified length, share of various parts of speech (POS), punctuation marks, co-occurrence of POS etc.

N-gram method was modified. Firstly, we have considered as a gram a word and not a symbol. We have examined the sequence of 3 elements in order to determine POS using Zalizniak's grammar dictionary. The main advantage of this method is resistance to synonymous substitution.

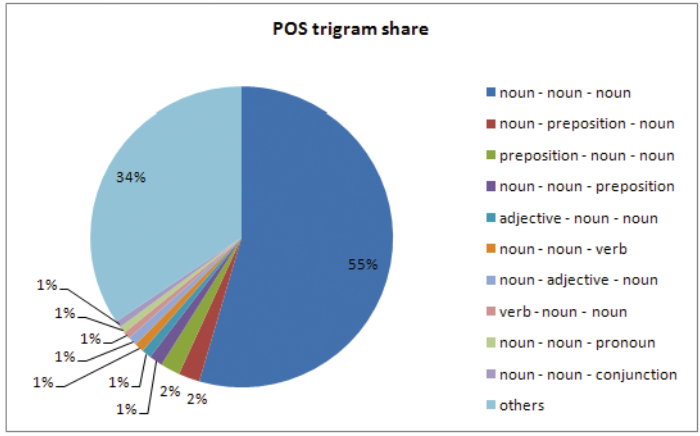

Figure 3. Share of various POS sequences

Spammers' tricks are detected by applying special rules. They are considered as spam indicators and they are used in the classifier. Moreover they allows to normalized distorted texts. In order to cope with deliberate word distortion we also use Damerau-Levenshtein distance. If a word is not the dictionary it is replaced by the closest one. However, this approach fails to deal with the majority of proper names, occasional words, some neologisms etc. because it tries to find the closest word which is not there.

Secondly, we have computed similarity of two messages:

$$similarity = \frac{2 \times NumberOfMatches}{(NumberOfTrigramsIn\_1text + NumberOfTrigramsIn\_2text)}$$

This quantity is not normalized. Similarity of Russian and Italian transforming messages is extremely high. Moreover, it varies slightly. Similarity of English and French letters is much smaller and has a large scatter (Fig. 4–Fig. 9).

It seems that N-gram approach is not efficient because words may be rearranged. However, even in natural languages with flexible word order (e.g. Russian) there are syntagmatic regularities. Deviations from these regularities perform an emphatic function or make a text difficult to understand. Perception difficulties are not desired in spam since they reduce the response.

As a classifier a support vector machine (SVM) has been chosen. The algorithm deals with retrieved quantitative characteristics taking into account possible distortion by applying rules and Damerau-Levenshtein

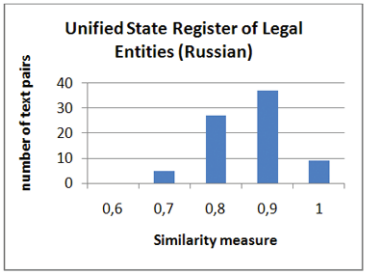

Figure 4. Trigram similarity measure of "ЕГРЮЛ"

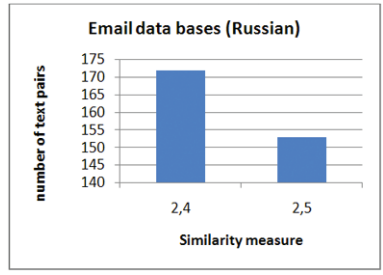

Figure 5. Trigram similarity measure of "Email базы"

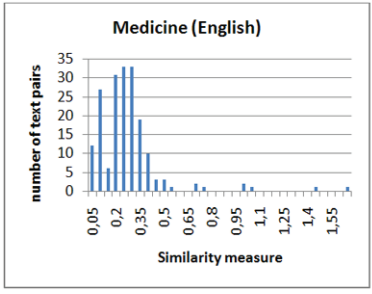

Figure 6. Trigram similarity measure of "Medicine"

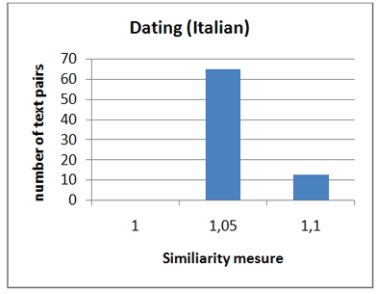

Figure 7. Trigram similarity measure of "Dating" mails

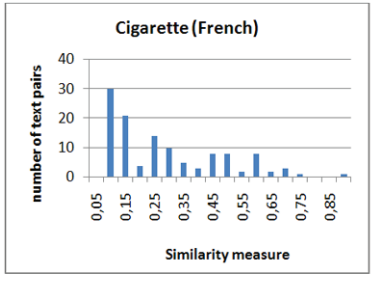

Figure 8. Trigram similarity measure of "Cigarettes"

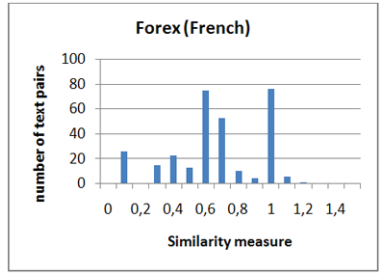

Figure 9. Trigram similarity measure of "Forex"

distance. We have used STATISTICA 8.0. Quantitative features of Russian messages enable to identify transforming messages with high accuracy. SVM parameters are given in the Table 1. As we can see SVM detects transformers with extremely high accuracy. However, the results obtained by SVM may be checked by N-gram method. It is possible to use other classifiers (e.g. neural networks are quite efficient).

Table 1. SVM parameters for the identification of Russian transforming messages

Sample size = 707 (Train), 236 (Test), 943 (Overall) Support Vector machine results: • SVM type: Classification type 1 (capacity=10,000) • Kernel type: Radial Basis Function (gamma=0,007) • Number of support vectors = 118 (0 bounded) Class. accuracy (%) = 100,000(Train), 100,000(Test), 100,000(Overall)

Thus, there are three main steps of transforming messages detection:

1. Quantitative features retrieval;
2. Classification using SVM;
3. N-gram verification.

6. CONCLUSION

Nowadays there are quite a lot of spam filters. Nevertheless, they are not efficient enough or they are very time and resource consuming. The majority of techniques are suitable only for email filtering. Incontrast to them content methods may be applied to spam filtering in various message systems (IM, social networks etc.). The improvement of signature methods seems to be topical. The proposed techniques enable to identify transforming messages in a very efficient way. 135 quantitative features have been retrieved. Almost all these features do not depend on the language. They underlie the first step of the algorithm based on support vector machine. The next stage is to test the obtained results applying N-gram approach. Special attention is paid to word distortion and text alteration. The main advantage of this method is resistance to synonymous substitution. In order to cope with deliberate word distortion we use Damerau-Levenshtein distance and substitution rules. If a word is not the dictionary it is replaced by the closest one. This methodisnotextremelyresource. It is portable to other languages (it needs only dictionaries). The ongoing work is being carried out in synonymous substitution analysis.

REFERENCES

1. Лаборатория Касперского Что такое спам Securelist, 2010 http://www.securelist.com/ru/encyclopedia/spam?chapter=151
2. **Наместникова, М.** Спам в декабре 2010 года Securelist, 2011 http://www.securelist.com/ru/analysis/208050676/Spam_v_dekabre_2010_goda
3. Лаборатория Касперского Спам в первом квартале 2010 года Лаборатория Касперского, 2010 http://www.kaspersky.ru/news?id=207733226
4. **Byun B., Chin-Hui Lee, Webb S.,Irani D., Pu C.** An Anti-spam Filter Combination Framework for Text-and-Image Emails Proceedings of the Sixth Conference on Email and Anti-Spam 2009

5. **Shi Pu, Cheng-Chung Tan and Jyh-Charn Liu** SA2PX: A Tool to Translate SpamAssassin Regular Expression Rules to POSIX Proceedings of the Sixth Conference on Email and Anti-Spam 2009
6. **Стахов, В.** Фильтрация спама при помощи системы rspamdБиблиотека Webcrunch, 2010 http://webcrunch.ru/library/administration/security/rspamd/
7. SpamAssassin - mail filter to identify spam OpenNet, 2004 [Cited:] http://www.opennet.ru/prog/info/1432.shtml
8. **Митрофанов, Е.** Антиспам «своими руками»: плюсы и минусы Securelist,http://www.securelist.com/ru/analysis/208050325/Antispam_svoimi_rukami_plyusy_i_minusy?print_mode=1
9. **Esquivel H., Mori T., Akella A.** RouterLevel Spam Filtering Using TCP Fingerprints:Architecture and Measurement-Based Evaluation Proceedings of the Sixth Conference on Email and Anti-Spam 2009
10. **Faynberg I., Hui-Lan Lu, Perlman R., Zeltsan Z.** Method and apparatus for reducing email spam and virus distribution in a communications network by authentication the origin of email mssages 7752440 USA, July 6, 2010
11. Лаборатория Касперского Эволюция спама securelist.com, 2009 http://www.securelist.com/ru/encyclopedia/spam?chapter=155
12. Яндекс Некоторые автоматические методы детектирования спама, доступные большим почтовым системам Компания Яндекс, 2010 http://company.yandex.ru/public/articles/antispam.xml
13. **Сегалович И., Тейблюм Д., Дилевский А.** Принципы и технические методы работы с незапрашиваемой корреспонденцией Яндекс, 2010 http://download.yandex.ru/company/spamooborona-latest.pdf
14. Лаборатория Касперского Электронный журнал «Спамтест», 2009 http://www.kaspersky.ru/news?id=143937135
15. **Broder, A.** On the resemblance and containment of documents Digital Systems Research Center, 2003 http://ftp.digital.com/pub/Digital/SRC/publications/broder/positano-final-wpnums.pdf
16. USENIX Conference Manber, U. 1994 Finding similar files in a large file system
17. **Sullivan D.** What Is Search Engine Spam? Search Engine Land, 10 21, 2008 [Cited: 05 01, 2011] http://searchengineland.com/what-is-search-engine-spam-the-video-edition-15202
18. **Gyöngyi Z., Garcia-Molina H.** Web spam taxonomy 2005
19. **Шарапов Р.В., Шарапова Е.В.** Алгоритм обнаружения ссылочного спама Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009» 2009
20. YandexСоветы вебмастеру Компания Яндекс, 1997-2011 http://help.yandex.ru/webmaster/?id=995298#995342
21. Лицензия на использование поисковой системы Яндекса Компания Яндекс, 1997-2011 http://company.yandex.ru/legal/termsofuse/
22. **Sew, S.** What Are Doorway Pages? Search Engine Watch, 2007 http://searchenginewatch.com/article/2048653/What-Are-Doorway-Pages

23. ANTULA. Виды поискового спама ANTULA, 2002-2011 http://www.antula.ru/rang-search_spam2.htm
24. **Ntoulas A., Manasse M.** Detecting spam web pages through content analysis In Proceedings of the World Wide Web conference 2006
25. **Mishne G., Carmel D. and Lempel R.** Blocking blog spam with language model disagreement In Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web 2005
26. **Urvoy T., Chauveau E., Filoche P.** Tracking Web Spam with HTML Style Similarities ACM Transactions on the Web 2006
27. **Павлов А.С., Добров Б.В.** Метод определения массово порождаемых неестественных текстов Компьютерная лингвистика и интеллектуальные технологии 2010
28. **Kleinberg, J.** Authoritative Sources in a Hyperlinked Environment Journal of the ACM 46, 1999, 5
29. **Page L., Brin S., Motwani R., Winograd T.** The PageRank citation ranking: Bringing order to the web. Technical report s.l. : Stanford University, 1998
30. **Chakrabarti, S.** Mining the Web: Discovering Knowledge from Hypertext Data 2003
31. **Coulthard, M.** Author Identification, Idiolect and Linguistic Uniqueness 2004
32. Linguistic Profiling for Author Recognition and Verification Halteren, Hans van 2004, Proceeding ACL '04 Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics
33. Идентификация языка UNICODE-текста по N-граммам длиной до 4-х включительно (квадрограммам) Сотник, С.Л. 2006, Математичне моделювання, pp. 111-114
34. N-Gram-Based Text Categorization Cavnar W. B., Trenkle J. M. 1994 Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval pp. 161-175
35. Идентификация авторства методами искусственного интеллекта Мещеряков Р.В., Васюков Н.С. 2005 Научная сессия ТУСУР
36. Авторский инвариант русских литературных текстов Фоменко В.П., Фоменко Т.Г. 1983 Методы качественного анализа текстов
37. **Рахимова, А.А.** Лингвистическая экспертиза Вестник КАСУ 2005
38. **Галяшина, Е. И.** Основы судебного речеведения 2003
39. **Coulthard, M.** Author Identification, Idiolect and Linguistic Uniqueness Applied Linguistics 2004, 25, 4